Successful Impact Evaluations: Lessons from DFID and 3ie

Edoardo Masset[1], Francis Rathinam,[2] Mega Nath[2], Marcella Vigneri[1], Benjamin Wood[2]

January 2019

[1] Centre of Excellence for Development Impact and Learning
[2] International Initiative for Impact Evaluation (3ie)

Suggested Citation: Masset E, Rathinam F, Nath M, Vigneri M, Wood B 2019 Successful Impact Evaluations: Lessons from DFID and 3ie. CEDIL Inception Paper No 6: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium initiative supported by UKAID through DFID. The mission of the centre is to develop and promote new impact evaluation methods in international development.

Corresponding Author: Edoardo Masset, email: edoardo.masset@lshtm.ac.uk

Table of Contents

Section 1 3

Introduction 3

Section 2 3

Background 3

SECTION 3 6

Methodology 6

SECTION 4 7

Conceptual framework 7

Design and planning 9

Implementation 11

Results 13

Cross-Cutting Themes 14

Section 5 16

Selection of the studies 16

Selection of DFID studies 16

Selection of 3ie Studies 17

Coding of the studies 18

Section 6 19

DFID Impact Evaluations 19

Successful DFID Evaluations 19

Characteristics associated with Success in DFID Evaluations 23

Section 5 25

3ie Impact Evaluations 25

Successful 3ie Studies 25

Characteristics of successful 3ie-funded evaluations 29

Unsuccessful Studies 31

Section 8 34

Conclusions 34

Section 9 35

Relevance for CEDIL 35

References 38

Abbreviations

3ie	International Initiative for Impact Evaluation
AGEP	Adolescent Girls Empowerment Programme, Zambia
CEDIL	Centre of Excellence for Development Impact and Learning
CGD	Child Development Grant
DAC	Development Assistance Committee
DFID	Department for International Development
DRHR	Delivering Reproductive Health Results
IEAINB	Impact Evaluation of Accelerated Improved Nutrition in Bangladesh
IRB	Institutional Review Board
IRC	International Rescue Committee
IV	Instrumental variables
ORCHSP	Operational Research Capacity in the Health Sector Project
RBAESE	Results-Based Aid in the Education Sector in Ethiopia
PSM	Propensity Score Matching
RCT	Randomised Controlled Trial
RFB	Result Based Financing
SAGE	Social Assistance Grants for Empowerment
SCG	Senior Citizens Grant
SSAPR	Security Sector Accountability and Police Reform Programme
VFG	Vulnerable Family Support Grant

Abstract

This paper explores the factors associated with the success of impact evaluations, drawing on the experience of two major funders of impact evaluations: the UK Department for International Development (DFID) and the International Initiative for Impact Evaluation (3ie). We first define successful evaluations along three dimensions: reliability, relevance, and policy influence. Based on a review of the literature we build a conceptual framework including the main factors associated with success at the stages of design, implementation, and analysis of results. We review selected impact evaluations recently funded by 3ie and DFID, we identify successful ones and we discuss characteristics associated with success. We find that careful planning, research team skills and composition, and sample size of the study are key elements of success. We also discuss unsuccessful cases of impact evaluations and suggest some guidance with regards to discontinuing studies facing implementation problems. We conclude with some reflections to help researchers and funders concentrate resources and efforts where they are most needed for the success of impact evaluations.

Section 1

Introduction

The goal of this paper is to identify factors that lead to successful impact evaluations of development interventions. Lessons from successes and failures can be used by researchers and funders to design and monitor better impact evaluation studies. It is well known that impact evaluations often do not achieve their intended aims, but how is success defined? The paper proposes a loose 'theory of change' of successful impact evaluations and identifies the potential factors determining success. We conduct a review of impact evaluations recently funded by DFID and 3ie to identify successful cases and factors associated with success. The goal of the paper is to offer a set of recommendations to improve the likelihood of success and avoid failures at the various stages of the evaluation process, including design, planning, and analysis of results. The paper consists of a review of the literature, a conceptualisation of success and associated factors, and a reflection on the relative importance of these factors. We are not able make causal claims regarding relationships between studies characteristics and success. Rather, we formulate some ideas and hypotheses that can become the subject of further investigation.

Section 2

Background

There is no shared or agreed definition of what constitutes an impact evaluation (Stern et al., 2012, Hearn and Buffardi, 2016). Development organisations define impact evaluation in different ways, but definitions share some common elements: impact evaluations assess social interventions, they assess changes in living standards, and they try to attribute the observed changes to the intervention. For the purposes of this paper, we narrowly define impact evaluations as studies that assess the impact of development interventions on welfare outcomes using counterfactual analysis through experimental or quasi-experimental designs.

We begin by defining 'success' of impact evaluations. Since impact evaluations have different goals and are conducted in different ways, any indicator of success must be multidimensional. Based on our reading of the literature, we identified three key dimensions of success:

- **Reliability:** a successful impact evaluation must be answering the evaluation questions in a credible and convincing way
- **Relevance:** a successful impact evaluation must provide evidence that is valuable and that can be used for making policy decisions or to inform policies
- **Use:** an impact evaluation is more successful when its recommendations are used by policy-makers

Studies have shown that credibility of results is vital to their use (Weiss and Buculavas, 1980 and Lavis, 2003). The belief that credibility is an essential element of policy impact is common among funders of impact evaluations (Gueron et al, 2013). However, authors have rarely defined credibility and rather stopped at describing those factors that make a study more or less credible to the eyes of the user (Jacobson et al, 2006). For the purpose of our work, we adopt a normative definition of credibility. In our paper 'reliable evidence' is the causal estimation of the effects of programmes, normally conducted using experimental and quasi-experimental methods. The standards employed to assess the quality of evidence are the object of debate and we adopt here those used by the Campbell Collaboration and the International Initiative for Impact Evaluation (3ie). These criteria are analogous to those used to qualify studies as 'impact evaluations' by the same organisations, which we also follow in this paper.[1] These criteria are defined based on methodologies that are considered able to infer causality in the evaluation of social programmes (see for example Shadish et al., 2002).

We define an evaluation as 'relevant' when it produces results that can be used to improve living standards of a population or group. A first element of relevance is therefore the production of useful results. Studies producing inconclusive results are not relevant. A second element of relevance is the importance of the evaluation question. For example, the Development Assistance Committee (DAC) evaluation criteria employ the term `relevance' to classify interventions that address priority development questions. The results of an intervention that affect a marginal fraction of the population, that cannot be scaled-up or extrapolated to other populations are not relevant. Similarly, results that come too late in the decision process or that are redundant, are not relevant, unless they are meant to inform decisions about other projects or to generate knowledge.

The results of an evaluation are unlikely to materialize or to be meaningful if the project being evaluated is not successfully implemented. Failure to participate in an intervention by its intended beneficiaries is the most common cause of failure (Rossi et al, 2003). A reasonable quality in the implementation of an intervention, and a sufficiently large level of participation in a project, are therefore pre-requisites for an evaluation study to be relevant.

Drawing on a large literature, we consider two fundamental uses of evaluation: instrumental and conceptual (see for example, Alkin et al, 2017, Cousins et al, 2002, and Nutley et al, 2008). We do not consider the symbolic use of evaluation – when studies are undertaken to fulfil a requirement or to legitimate decisions already been made – and other misuses of evaluation (McNulty, 2012). The instrumental use occurs when results are used to make contingent decisions, for example, about discontinuing, modifying, or scaling-up an intervention. The conceptual use occurs when a study

[1] See for example the criteria used by 3ie for inclusion in the 3ie Repository of Impact Evaluations, which includes all impact evaluations of developed interventions (4,260 as of November 2017) that have been published through 2017 (Mishra & Cameron, 2014). The repository includes studies that employ any of the following methodologies: randomized controlled trial (RCT), regression discontinuity design (RDD), propensity score matching or other matching methods, instrumental variable (IV) estimation (or other methods using an instrumental variable such as the Heckman Two Step approach), difference-in-differences (DD), or a fixed or random effects model with an interaction term between time and intervention for baseline and follow-up observations

changes policy-makers' perspectives about a specific issue or causal mechanisms, sometimes in the long term and in a subtle way. We abstract from the 'process' use of evaluation, whereby studies improve the skills of project managers and other stakeholders and we also abstract from models of 'evaluation influence' (Herbert, 2014). Briceno et al. (2011) provide some examples of process use in international development, while Legovini et al. (2015) analyse similar 'spill-over' effects of World Bank evaluations.[2] Process and spillover effects are important but also difficult to track empirically.

A final element of success is value for money. Impact evaluations can be very expensive and well-funded studies are more likely to succeed. For example, MEASURE Evaluation presents a range of project costs between $400.000 and $850.000, with a single study costing over $3.5 million (MEASURE evaluation, 2015). 3ie estimated an average cost of $336.000 for impact evaluations funded between 2009 and 2015 (Puri & Rathinam, 2017). It is believed that impact evaluations are cost-effective in the long run because they improve project design and inform global portfolios (Stickler, 2016), but little empirical evidence is available in support of this hypothesis. In principle, we would rate an evaluation more highly when it achieves the same results of a similar study but using fewer resources.

There is a considerable literature exploring factors affecting the use of evaluations. Cousins et al. (1986) identify the following characteristics of successful use: relevance, credibility, evaluation quality, communication quality, findings and timeliness, information needs, decision characteristics, political climate, competing information, personal characteristics, and commitment/receptiveness to evaluation. Johnson et al. (2009) add two additional characteristics to the list: stakeholders' engagement and evaluators' competence. These authors build their conceptual frameworks as 'loose' theories of change (Davies, 2016). In a 'loose' theory of change the goals of an intervention are defined but the determinants and the causal mechanisms are not. Our paper follows a similar approach. First, we build a loose theory of change of the factors associated with the design, implementation and use of evaluations. Second, we analyse the associations between successful evaluations and characteristics.

The remainder of the paper is structured in the following way. Section 3 outlines the methodology of our study. Section 4 builds a conceptual framework of successful impact evaluations and associated factors. Section 5 discusses the characteristics associated with successful evaluations funded by DFID while section 6 covers characteristics associated with successful studies funded by 3ie. Section 7 presents the conclusions of the study, while section 8 discusses how these inform the programmes of work promoted by CEDIL.

[2] Legovini et al (2015) show that World Bank projects, conducted alongside impact evaluations, are more likely to be successful because they tend to be implemented as planned and therefore more likely to achieve their stated goals. The authors speculatively attribute this effect to (a) better planning and evidence-base in project design, (b) greater implementation capacity due to training and support by research team and field staff, (c) better data for policy decisions, and (d) observer effects and motivation.

SECTION 3

Methodology

The goal of this paper is, first, indentifing characteristics associated with the success of evaluations and, secondly, establishing the strenght of the associations between characteristics and success in order to inform evaluation practice. As noted in the background section, there is a considerable theoretical work in this area but much less empirical work.

We begin by reviewing the literature on the characteristics of successful studies. In doing so we rely on our expert knowledge of the literature, on a search of key journals, complemented by a search of blog pages. We screened all titles and abstracts of the *Journal of Development Effectiveness* (from 2009) and of *Evidence and Policy* (from 2005). We used Boolean search using the string 'credibility relevance use success failure' to search relevant titles published since 2005 in the *American Journal of Evaluation*, in the *Evaluation Review*, and in *Evaluation*. Impact evaluations in international development are relatively new and much discussion on characteristics of successful evaluations has occurred in blogs rather than in peer-reviewed journals.[3] We conducted a search of selected blogs. We used our expert knowledge to select 14 blog pages on evaluation and international development.[4] Few blog pages allow Boolean search and in most cases we searched for 'impact evaluation' and retrieved all the blogs available. We found 16 entries that were highly relevant to our study. Blogs sometimes included references to papers, reports and books, seven of which became part of our review.

After identifying the main factors thought to be associated with successful evaluations we build a 'theory of change' of successful studies. In doing so we use a 'loose' theory of change approach (Davies, 2016), whereby factors affecting success are represented diagrammatically at each stage of the evaluation process, but without spelling out the underlyng causal links and assumptions becasuse these are not known with precision.

After building a conceptual framework for the analysis we conduct our empirical work on samples of impact evaluation studies funded by DFID and by 3ie. We identify all impact evaluations funded by 3ie and DIFID within a pre-specified period of time, we classify each study as successful/unsuccesful, and we code all available characteristics

[3] The literature search identified 52 papers based on title and abstract distributed in the following way: *Journal of Development Effectiveness* (22), *American Journal of Evaluation* (19), *Evaluation Review* (4), *Evaluation* (3) and *Evidence and Policy* (4). Of these, 22 papers were eventually included in our review.

[4] The selected blog pages are: *A tip a day by and for evaluators of the American Evaluation Association* (http://aea365.org/blog/tag/blog/); the *GiveWell* blog (https://blog.givewell.org/category/impact-evaluation/); *Sharing information to improve evaluation* of Better Evaluation (http://www.betterevaluation.org/blog); *What works: transforming development through* evaluation of the Independent Evaluation Group at the World Bank (http://ieg.worldbankgroup.org/blogs); *Innovations for Poverty Action* (https://www.poverty-action.org/blog); *Evidence matters* of the International Initiatives for Impact Evaluation (http://blogs.3ieimpact.org/); *Development impact* of the World Bank (https://blogs.worldbank.org/impactevaluations/blog); the World Bank's web page, *Failure* (http://blogs.worldbank.org/impactevaluations/failure). *R&E search for evidence* of FHI360 (https://researchforevidence.fhi360.org/); *Evaluate*, the MEASURE evaluation blog (https://measureevaluation.wordpress.com/); impact blogs of USAID (https://blog.usaid.gov/); *Views from the centre* of the Centre for Global Development (https://www.cgdev.org/global-development); *Evidence for action* of UNICEF (https://blogs.unicef.org/evidence-for-action/); *Shaping policy for development* of ODI (https://www.odi.org/comment).

associated with that particular study.[5] Search and coding of 3ie and DFID studies were conducted by different teams operating independently. Two 3ie researchers selected, screened and coded the 3ie studies. One researcher identified the DFID impact evaluations and two researchers independently screened and coded the same studies. Studies were coded along a number of factors related to success and discussed in Section 4 and 5. 3ie and DFID studies were coded in different ways because of the different level of information available. In the case of 3ie studies we were able to access evaluation reports, as well as internal and external reviews, and we were able to build a large database of evaluation characteristics. In the case of DFID evaluations, we relied on much more limited information and we were able to code only few of the the relevant characteristics.

Since we were not able to analyse 3ie and DFID studies in the same way, the results of the assessments are not comparable. Many aspects of the evaluation design and implementation could not be assessed for DFID studies. This is particularly obvious in the case of indicators of policy impact. 3ie routinely collects data from authors and users to assess the potential uses of evaluations. No similar data is available for DFID studies and we had to rely on management responses. Management responses often reflect only the instrumental use of evaluations thus neglecting its conceptual use.

Since the number of studies selected was small in relation to the number of associated characteristics, we are unable to use regresssion analysis and statistical testing to assess the validity and the strength of the associations. Therefore we present the results using simple tabulations and averages along with a narrative synthesis supplemented by case studies of positive and negative deviants. We define as positive deviants those evaluations that are not reliable and that are nevertheless successful in influencing policy making. Conversely, negative deviants are evaluations that are reliable but that do not have an impact on policy.

SECTION 4

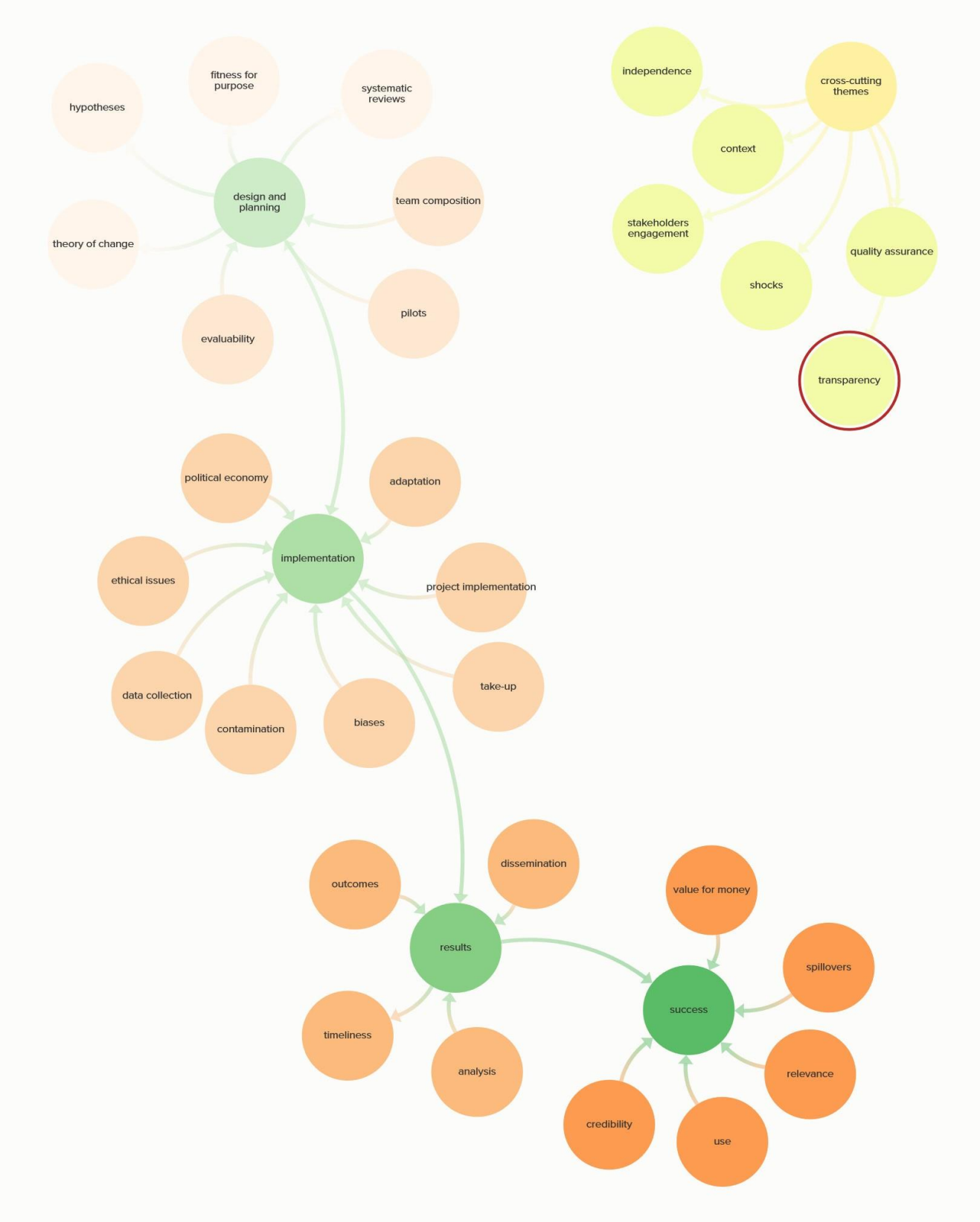
Conceptual framework

One way of building a theory of change of successful evaluations would be describing an 'ideal' impact evaluation. That is, an impact evaluation having all the ingredients, and in the right combination, to be successful. What makes an evaluation successful, however, depends on the context and on the goals of the exercise. We therefore opted for a different approach. We first defined the steps required to complete an impact evaluation and we then identified the factors that affect the study, positively or negatively, at each step. We considered the following stages of an evaluation: design, implementation, and results. We also include in the framework some cross-cutting factors that do not belong to any specific stage. The resulting framework is a 'loose' theory of change (Davies, 2016 and Davies, 2018), which associates characteristics to outcomes without spelling out causal links and underlying assumptions because these

[5] DFID definition of impact evaluation is broader than 3ie's definitions and includes studies that are not experimental or quasi-experimental. In our study however we only considered impact evaluations studies as defined by 3ie.

are unknown or too heterogeneous across different contexts (see Figure 1). In what follows we briefly describe all relevant characteristics identified at each stage.

Figure 1 Factors Affecting 'Success' of Impact Evaluations



Note: Figure was produced using the Kumu software (https://kumu.io)

Design and planning

Clarity of hypotheses

The first step in designing an impact evaluation is defining the evaluation questions and related hypotheses. But where are evaluation questions coming from? From a researcher's perspective questions originate in social and economic theory or from research and evaluation experiences in other countries or contexts. In these cases, researchers look for opportunities to test these hypotheses in the field (Karlan & Appel, 2016a). In some other cases, evaluators assess the impact of established interventions. This does not mean that researchers cannot test modifications of the original interventions. Often the interventions are not fully spelled out and there is room for the researchers to tinker and adjust the interventions (Duflo, 2017). Sometimes, evaluation questions are formulated by decision-makers wanting to learn about the impact or the scalability of interventions and sometimes questions simply relate to the extension or termination of a project for accountability purposes.

The goals of evaluations and the type of hypotheses formulated have implications for the relevance and use of the results. For example, a study motivated by accountability may teach very little about what works but may lead to immediate policy decisions. On the other hand, an innovation study may reveal impacts that were previously unknown but without affecting policy. The objectives and the scope of the evaluation should be clearly defined, preferably with the support of detailed terms of references, a theory of how the intervention is expected to work, and a review of the effectiveness of previous interventions.

Theories of change

It is well accepted that impact evaluations should be based on a theory of how interventions are expected to affect the outcomes (White, 2009). An evaluation is more valuable if it not only explains whether a project worked or not but also why and how. Development agencies advocate the use of theories of change in the evaluation of social interventions. Though theory of change representations have become particularly popular, structural models, path diagrams and similar diagrammatic devices serve the same purpose of illustrating the causal links between final and intermediate outcomes, and their determinants.

Systematic reviews

Systematic reviews summarising the available evidence on the effectiveness of interventions and other synthesis products such as evidence gap maps, meta-analyses and critical reviews, are an obvious starting point of every impact evaluation. Reviews often include conceptual frameworks of interventions and detailed descriptions of challenges, biases and other issues arising in the collection and analysis of the data. The Cochrane Handbook of systematic reviews states that conclusions of systematic reviews should inform researchers of the need and the nature of future research on the same topic (Schünemann et al., 2011).

Fitness for purpose

A study can be very well designed and executed but may not be the right one to answer a specific evaluation question. For example, a randomised control trial might be appropriate to test the effectiveness of a new service but might be silent on the reasons of failure and success unless conducted alongside other types of evaluation. Studies whose methods and purposes are misaligned are not 'fit-for-purpose' (Gough, 2007). These types of studies are methodologically sound but do not necessarily provide answers to questions that are relevant to policy-makers.

Evaluability

Ideally, an evaluability exercise should be conducted to assess the technical, political and practical feasibility of an evaluation before it starts. The OECD-DAC defines evaluability as 'the extent to which an activity or project can be evaluated in a reliable and credible fashion'. Davies (2015), argues that an evaluability assessment reduces the risk of irrelevant or invalid findings, thus preventing the waste of resources on inappropriate evaluations. Evaluability assessments of impact evaluations should aim at answering three broad questions: a) is it plausible to expect impact? b) would an impact evaluation be useful and used? and c) is it feasible to assess or measure impact? Peersman et al. (2015) provide guidelines, and a useful checklist, to assess the evaluability of impact evaluations. Evaluability assessments are particularly needed when using rigorous evaluation methods like random assignment because some interventions or contexts are not amenable to this type of evaluation (Wholey, 2004, Epstein et al., 2012, and Greenberg et al, 2014). Sometimes the evaluability of an intervention needs to be probed through a pilot evaluation.

Pilots

Many interventions fail because cannot be implemented in the planned way. These problems affect in particular new interventions that are tested for the first time. In some cases, interventions are implemented in the wrong settings. Many such examples are reported by Karlan (2016a) and summarised in the following way by Mckenzie (2016): '*doing projects in the wrong place* (e.g. a malaria prevention program where malaria is not such an issue), *at the wrong time* (e.g. when a project delay meant the Indian monsoon season started, making roads impassable and making it more difficult for clients to raise chickens), *with a technically infeasible solution* (e.g. trying to deliver multimedia financial literacy in rural Peru using DVDs when loan officers were not able to find audio and video setups).' Piloting an intervention (also called 'formative evaluation' or 'field testing') may overcome this problem (White, 2014). Pilots also provide an idea of project take-up and effect sizes thus helping with power calculations and the definition of the sample size (Wood and Djimeu, 2014).

Team's composition

Evaluation teams need the right set of technical skills and the right combination of disciplinary approaches. 'Hiring the wrong person(s) on the impact evaluation team, or not defining their responsibilities well' was noted as the second most common mistake in conducting impact evaluations among World Bank staff (Vermeersch, 2012). Typically, impact evaluations require skills in qualitative and quantitative research methods.

Evaluation of social interventions deal with behavioural change and teams need to include experts in psychology, sociology, economics and in the specific subject matter. Since the ultimate goal of evaluation is informing policy, teams should also include members with knowledge management skills to disseminate findings and for engaging with stakeholders. Team composition and experience are commonly used as screening and scoring criteria by agencies commissioning impact evaluations.

Implementation

Study implementation

Evaluations are rarely implemented in the planned way and in some cases practical problems unforeseen at the design stage derail implementation. The problem is more serious in experimental designs in which the design of the project and the design of the evaluation are one and the same. Mckenzie (2016a) provides two such examples of World Bank projects aiming at bringing enterprises in the formal sector that went wrong. In Brazil, new iPads to be used as prizes had to be returned to the US for tax reasons. In Peru, firm's distance to the tax office could not be used as an instrument for the estimation of project effects because tax collectors, in order to reach their quotas, would simply inspect the closest firms they could find.

Project Take-Up

Low take-up has been indicated as the single most common reason of project failure (Rossi et al., 2004). Evaluations often assume that all individuals in a project area are beneficiaries and data on take-up or participation in the interventions are rarely collected. The assumption that all people in a given area join the interventions is highly unrealistic and studies that collect data on people's participation in interventions find that they progressively lose out their intended beneficiaries through a 'funnel of attrition' (White, 2014). For example, IPA partnered with Global Giving on a program that invited clients of local organisations to provide feedback about the services received by sending SMS messages to Global Giving. The client feedback would in turn be sent to potential donors, and the researchers were hoping to find whether feedback would change service provision and donors' attitudes. A total of 37 organisations in Guatemala, Peru, and Ecuador were identified and randomised in the study, but only 5 took part (Karlan & Appel, 2016b). In another example (Mckenzie, 2016a), a team spent $40,000 to send letters to financial institutions in Mexico to offer financial education. The team was hoping to obtain between 800-1200 responses out of the 40,000 letters sent, but in the end they obtained only 42 replies. Similar examples of low project take-up abound in the literature.

Contamination of the Control Group

Even the best designed evaluation may be spoiled by contamination of the control group. This occurs when the services offered by the intervention are accessed by residents in the control areas. In some cases, contamination can be irreparable. For example, a randomised evaluation of an HIV intervention in Andhra Pradesh was aborted after the funder, in collaboration with government authorities, decided to

saturate the state with a nearly identical and much larger intervention that wiped out the randomised control group (Samuels & McPherson, 2010).

Biases in evaluation

Several biases originating at the implementation stage may affect the results of an impact evaluation. Biases include high attrition of study participants, non-compliance, placebo and Hawthorne effects with inability to blind.[6] These types of issues are widely discussed in the evaluation literature (Shadish et al., 2002). A non-technical discussion of many of these problems, particularly in relation to randomised controlled trials, can be found in Glennerster et al. (2013) and Greenberg et al. (2014). Quasi-experimental studies are even more vulnerable to biases and to the inherent difficulty of controlling for unobserved differences between the comparison groups.

Sample Size and Data Quality

Most development interventions produce small effects that need a large sample size in order to be detected at the desired level of statistical significance. Wood and Djimeu (2014) discuss how rarely evaluators conduct and report power calculations for determining sample size. As a result, evaluations often have sample sizes that have inadequate statistical power (White, 2014), particularly for outcomes, like income and expenditure, that are measured with error. The problem is compounded if intra-cluster correlations are not taken into account in the power calculations.

Even if researchers get the sample size right, many things can go wrong during data collection. 'Issues in the design and (preparation of) data collection' were by far the most represented category in a survey of mistakes in impact evaluations among World Bank staff (Vermeersch, 2012). Problems included: forgetting individual identifiers, or not printing the cover page of the questionnaire, and missing information on the respondent. Attrition, the process of losing over time the units of observation originally selected for the interviews, is another common problem, particularly in areas and at times when populations are more mobile, such as in slums and in conflict areas. Attrition, though often ignored by researchers, compromises the representativeness of the results and the unbiasedness of the estimates (Greenberg and Barnow, 2014).

Ethical Issues

Research on human subjects must meet a minimum set of ethical standards. This may require approval from an Institutional Review Board, obtainment of informed consent, protection of confidentiality of the information and any other measure that prevents the researchers from doing any harm (Glennerster, 2017). The ethical cost of an evaluation implicitly sets a benchmark for its desirable success. It could be argued that, at a minimum, the benefit of an impact evaluation should outweigh its ethical cost.

[6] A placebo is any treatment prescribed or self-prescribed, knowingly or unknowingly, for its therapeutic effects, but which actually is ineffective or not specifically effective for the condition being treated. A placebo effect is the non-specific therapeutic effect produced by a placebo (Shapiro & Shapiro, 1999). Hawthorn effects occur when project beneficiaries simulate project effects in order to achieve other goals. For example, subjects may pretend to be poor in order to obtain more benefits.

Processes of ethical clearance are notoriously slow and a frequent cause of evaluation delays.

Political Economy

Governments and implementers can be opposed to evaluations or to particular aspects of their design. Campos et al. (2012) provide a number of examples of failed experimental subsidy programmes for private firms, in which government officers were openly opposed to randomisation. The opposition was sometimes accentuated by turnover of government staff, which caused severe delays.

Ability to Adapt

Evaluations rarely follow pre-defined plans. Changes in circumstances require sudden adjustments in evaluation design or implementation. So much so that contingency plans are advisable, particularly when conducting randomised trials (Briceno et al., 2011). 'Creative adaptation of the evaluation design to fit operations is the norm not the outlier (Priedeman Skiles et al., 2014).' A review of 10 evaluations of health interventions by MEASURE found that each case faced design and implementation challenges that required creative solutions. These challenges included the identification and selection of programme beneficiaries, random assignment, identification of a control group, heterogeneity of impacts, timing of baseline data collection, and absence of baseline data (Priedeman Skiles et al., 2014). Puri and Rathinam (2017) discuss how the completion of impact evaluations funded by 3ie depended on their ability to adapt to unforeseen circumstances.

Results

Analysis

An impact evaluation must provide the best possible estimates of project impact. What makes a technically high-quality evaluation is the object of debate but there are common standards, such as, for example, risk of bias tools and assessments used in systematic reviews, and there are quality markers, like publications in peer-reviewed journals. Shadish et al. (2002) is a classical reference on the pros and cons of experimental and quasi-experimental designs.

Dissemination

The results of a study have to be disseminated in order to reach the desired audiences. Traditional dissemination methods include policy briefs, workshops, and publications. This type of communication can be effective in a context of rational decision making, but its role is limited in contexts that are politicised or that include multiple stakeholders. More recently, *knowledge brokers* have been charged with the role of disseminating evidence among policy-makers (Brown et al., 2018). However, the evidence on what works in research dissemination is limited and rigorous empirical studies of the policy impact of communication strategies are lacking (Nutley et al., 2008).

Outcomes

The selection of meaningful outcome indicators and its appropriate measurement and monitoring is a key element of impact evaluation design, but problems arise at the

analysis stage as well. Some evaluations present results on multiple outcomes without a pre-specified analysis plan, in such a way that results are inconclusive or not credible. Other evaluations summarise multiple outcomes with index indicators that have little practical meaning. A successful evaluation employs a transparent methodology to analyse and interpret the results. The analysis of multiple outcomes increases the possibility of purposely selecting statistically significant results and of finding impacts that are the result of chance (Gelman, 2013). Transparency in conducting research and statistical methods, like corrections for multiple testing of hypotheses, may help to address this problem.

Timeliness

The findings of evaluations should be reported as soon as they are available in order to support decisions to modifying or discontinuing programmes (Briceno, 2011). On the other hand, the effects of behavioural change interventions occur in the long term and the compression of timeframes to satisfy the policy cycle can compromise the ability to detect project effects by cutting down essential stages, such as pilots and consultations (de Silva, 2015). There is an unresolved tension between assessing welfare impacts that can only be observed after several years of an intervention, and the provision of evidence for making decisions, which follow particular political or financial cycles. Timely findings are more likely to be produced by monitoring activities, or by evaluations specifically designed to improve projects while they are operating, rather than by impact evaluations as a recent literature on adaptive management suggests (Pritchett et al, 2013).

Cross-Cutting Themes

This section includes factors affecting evaluations that, as depicted in Figure 1, are not specific to any evaluation stage (design, implementation, and result) but that are relevant at every stage throughout the entire evaluation process.

Stakeholder Engagement

Stakeholder engagement is by far the factor most commonly quoted in blogs[7] and in the literature (see for example Sturdy et al, 2014, Culbertson et al, 2014, Briceno et al, 2011). Project implementers, funders and users should be involved from the design stage of an evaluation. An intense collaboration between researchers and the implementing organisation is even more important when conducting a randomised trial because the researchers become directly involved in project design and implementation. Glennerster (2017) discusses the following elements of a fruitful collaboration between evaluators and managers when implementing a randomised trial: flexibility about evaluation design, sharing of expertise, providing intermediate products, and having a local presence.

Stakeholder engagement has become a mainstream concept in evaluation but there is no empirical research on its contribution to the success and use of impact evaluations

[7] See for example Stickler (2016), Keller and Savedoff (2016), Lopez-Avila (2016), Ferguson (2017), White (2014), Mckenzie (2016), MCC (2012), and Priedeman et al. (2014).

(Fleischer et al., 2009). Oliver et al. (2018) critically review the literature on stakeholder engagement and conclude that effective engagement with research findings include: facilitating access to evidence, building decision makers' skills, and fostering changes to decision-making processes.

Transparency

An evaluation is more likely to follow best practices when activities are open to the scrutiny of the public and of other researchers. Examples of ways to increase evaluation transparency include the public registration of the evaluation, and the publication of a pre-analysis plan, which set out the outcomes and the methodology of the evaluation. Evaluation designs should be ideally registered in order to avoid subsequent cherry-picking of results (White, 2014). Some authors have designed software to declare the study design in computer code form. Given the code declaration, the software can diagnose statistical power, bias, expected mean and external validity in advance (Blair, Cooper, Coppock, & Humphreys, 2016). Transparency also includes sharing of the datasets and the codes used in the analysis in order to allow replication of research (Wood, 2015).

Shocks and External Events

Project implementation and evaluations can be derailed by unexpected shocks. For example, a World Bank team conducted a survey of 2,500 households and 2,500 enterprises in rural Egypt to assess the impact of an intervention expanding financial access. But the social instability in the wake of the Egyptian revolution meant microfinance organisations did not want to expand activities to poor areas, and the intervention never took place (D. Mckenzie, 2016b). A risk assessment at the planning stage and the formulation of contingency plans might be helpful particularly when researchers set out to work in areas prone to conflict or natural disasters.

Implementation Context

The complexity of interventions, consisting of multiple interacting activities and projects at varying levels (countries, regions etc.) and including diverse stakeholders groups, is a challenge to the design of causal pathways (Vaessen, 2017) and to the identification of a control group (Cuong, 2014). This is often addressed by focusing on a narrow component of the programme rather than on the entire programme, thus offering little guidance in terms of scaling up the whole intervention (Lopez Avila, 2016). Successes and failures vary with the characteristics of the area in which the project is implemented, with the type of intervention and with the characteristics of the implementer. Implementer characteristics are particularly relevant for projects implemented by NGOs but that require governments for countrywide scale-up (Bold et al. 2013).

Quality Assurance and peer-review groups

Progress reviews during the evaluation and peer-review of the results at various stages aim to ensure the quality of the design, implementation and analysis. For example, Briceno (2011) states that the work of evaluators should be supervised through advisory

panels, peer review groups, and funders in order to ensure the technical quality of the research.

Independence

Independence is a fundamental element of evaluation. Evaluators should be independent of the programme managers and funders to ensure the credibility of the study (Briceno, 2011). However, too much independence can be harmful if it prevents a common understanding of the programme and of its evaluation (Sturdy, 2014). The argument in favour of independence might have been overstated (White, 2014). Independence may prevent access to key information and reduce the policy impact of the evaluation.

Section 5

Selection of the studies

The remainder of the paper assesses the relevance of the characteristics associated with success identified in the previous section. In this exercise we use a sample of studies supported by two major funders of impact evaluations: DFID and 3ie. Here, we briefly describe the selection of the sample of studies and the methods used to identify success and associated factors.

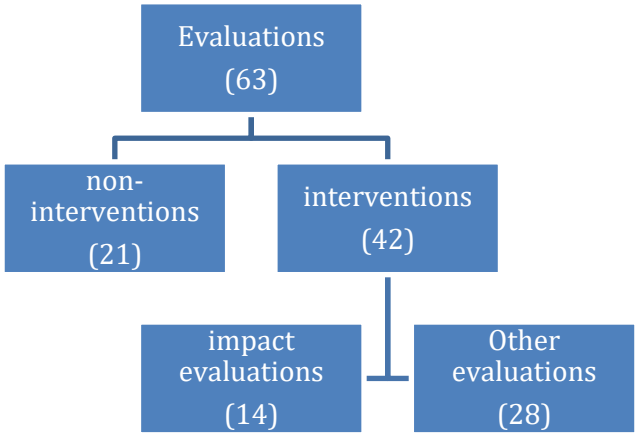
Selection of DFID studies

Every year DFID publishes a collection of commissioned evaluations in international development. The body of this work spans a twenty-year period from 1997 to 2017, and includes independent reports categorised by sector and country, together with a short document with DFID's management response. We examined all the evaluations commissioned by DFID since 2012, which included 63 studies. This set was subsequently complemented with impact evaluations published during the first trimester of 2018.

We employed two criteria to select studies for this review. These are the same criteria used by 3ie to incorporate studies in the 3ie database of impact evaluations (Mishra and Cameron, 2014). We adopted this protocol because it ensured consistency in the type of evaluations supported by 3ie and DFID considered in our study. The first criterion required that the study evaluated a development intervention. Following the 3ie protocol, we included studies that examined the effects of a specific policy, programme or intervention, and excluded 'studies which only investigate natural or market-based occurrences, or that report on the findings of controlled laboratory experiments with no discernible development intervention.' This led to the exclusion of 21 studies including, for example, *the Evaluation of DFID Online Research Portals and Repositories*, *Evaluation of the DFID-AFDB Technical Cooperation Agreement (TCA)*, and *Evaluation of the Consultative Group to Assist to the poor - Phase IV Mid-Term Report*. The second selection criterion was the use of a counterfactual method. We included studies that used an experimental or quasi-experimental design. This led to the exclusion of 7 desk-based reviews, 5 process

evaluations, 4 'theory-based' evaluations, 4 value for money cost analyses, 3 qualitative studies, 2 formative evaluations and 1 rapid health facility assessment.[8] The selection process (see Figure 2) delivered 14 evaluations for review.

Figure 2 Selection of DFID Impact Evaluations



Most impact evaluations were quasi-experimental (11), often consisting of difference-in-difference studies reinforced by matching methods or, in some cases, simple project-control comparisons at one point in time after the intervention. Three evaluations were experiments. The majority of impact evaluations were commissioned in the agriculture, health and population, and education sector. Some sectors like climate change and conflict management did not have a single impact evaluation while most evaluations of public sector management were not impact evaluations. All impact evaluations were conducted in South Asia and Sub-Saharan Africa and a larger proportion was conducted in South Asia.

Selection of 3ie Studies

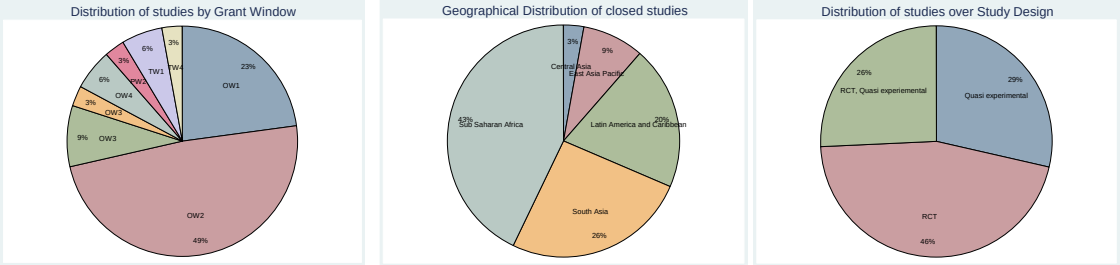
As of the end of 2017, 3ie had posted 112 impact evaluations on its website. Due to time and resource limitations, we drew a subsample of 35 studies for our analysis. To ensure a diverse representation of the 3ie-funded studies, we stratified the studies by thematic sector, geographic area, funding modality and technical quality. To obtain a diverse representation of sectors, we included all the sectors that had only one study and we limited the maximum number of studies from a specific sector to five.

3ie funding modalities include open windows, thematic windows and policy windows. Open windows allowed researchers to propose topics for evaluations while thematic windows aimed to increase the body of rigorous evidence in a specific sector. Policy windows supported evaluation of government programmes chosen by 3ie developing country members. Since closed grants were from open windows, about 88% of the

[8] We did not include in our review the experimental studies conducted under the Girls Education Challenge Fund because the available documents reported summary results of multiple impact evaluations without the level of detail needed in our review.

sampled studies were from open windows as well (see Figure 3). More than half of the studies were conducted in Sub-Saharan Africa and South Asia (see Figure 3). A third of the studies were quasi-experimental while all the remaining ones were experimental or a combination of experiments and quasi-experiments.

Figure 3: Selected 3ie evaluations by thematic window, geographic area and study design



Coding of the studies

We coded the 3ie and DFID studies along factors potentially associated with success (see Table 4). Some characteristics were coded numerically, for example, whether the study was reviewed by an Institutional Review Board (IRB) or sample size. Other characteristics consisted of qualitative judgments like, for example, the clarity of the conclusions or their policy relevance. We were able to extract information on most characteristics for the studies funded by 3ie but only limited information for the studies funded by DFID.

Table 4: List of factors associated with success and their proxy indicators extracted from the studies

Stage	Characteristic	Metric
Designing and planning	Hypotheses	Qualitative assessment of the statement of the problem the project and the evaluation set out to solve.
	Theory of change	A well-defined theory of change
	Reviews	Are systematic reviews quoted in reference to the problem statement?
	Team composition	Skill set of evaluators
	Evaluability	Evaluability assessment conducted
	Pilot	Was the intervention piloted?
	Fitness for purpose	Is the method appropriate to answer the evaluation question?
Implementation	Project implementation	Implementation issues and delays
	Take-up	Participation rates
	Biases	Attrition rates
	Contamination	Contamination reported
	Data collection	Sample size
	Ethical issues	IRB approval
	Political economy	No data available
	Adaptation	Letters of variation and modifications to study design
Results	Analysis	Sub-group analysis Intermediate outcomes

	Timeliness	Delays in reporting
	Outcomes	Effect sizes are reported and discussed, clarity of conclusion with respect to evaluation questions
	Dissemination	Policy influence plans, policy briefs, presentations and conferences, media, and publications
Success	Credibility	The study is rigorous and unbiased
	Relevance	Results are conclusive and meaningful to an improvement of living standards in a population or group
	Use	The study has a policy impact as assessed by our reading of the management response (in the case of DFID studies) and by 3ie's own appraisal (in the case of 3ie studies)
	Value for money	Budget size
Cross-cutting factors	Quality assurance	Peer review groups and advisory panel
	Context	Country and region, sector of intervention (World Bank classification)
	Shocks	External events affecting the evaluation
	Stakeholder engagement	Meetings and workshops with implementers, training events, and endorsement by the implementers
	Transparency	Study registration and pre-analysis plan
	Independence	Evaluation team independent of management and funders

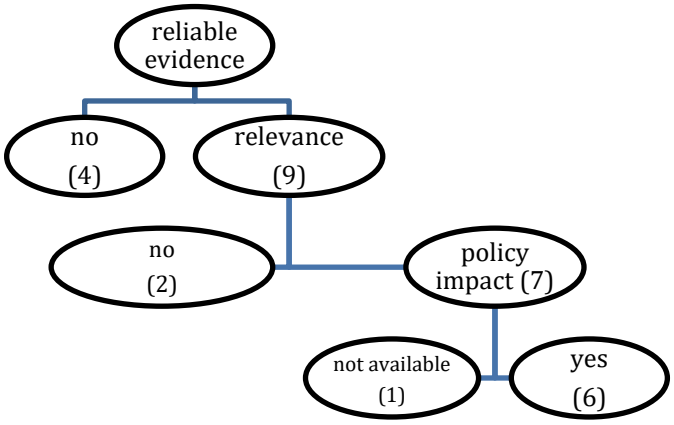
Section 6

DFID Impact Evaluations

Successful DFID Evaluations

We assessed the success of evaluations along three dimensions: reliability, relevance and policy impact. We assessed the reliability of the studies against the impact evaluation methods used. We classified studies as relevant based on our reading of the study conclusions and of the management response. We identified studies having policy impact by reading the recommendations in the reports and in the management response. Two researchers independently scored the studies and divergences of judgment were discussed and resolved. Figure 6 illustrates the classification of the studies.

Figure 6: Reliability, Relevance and Policy Impact (DFID Evaluations)



Note: as discussed in the text one of the studies was classified as not having policy impact because no information was available on impact

Four evaluations did not provide reliable evidence. One study assessed the impact of a capacity building intervention in several low income countries. The evaluation included a difference-in-difference analysis of the impact of the intervention on knowledge and skills of health care workers. However, the sampling methodology as well as the analysis were rather inaccurate. The quantitative assessment was only a minor component of the overall evaluation. The authors did not draw strong conclusions from the evaluation and management did not seem to take relevant lessons from this particular aspect of the evaluation.

An evaluation of a results-based education project in Ethiopia used an interrupted time-series design. The intervention had been implemented at the national level and no other design option was available. However, there were delays in project implementation and the data were not fully representative. This was acknowledged by the researchers, while the management response stressed that in this case lack of evidence of impact did not imply lack of impact.

The evaluation of a livelihood programme in Bangladesh found significant impacts on poverty reduction. However, the management response exposed the fact that external reviewers found large discrepancies between the reported income data and data collected by other agencies, which called in to question the credibility of the entire survey.

The fourth study was an impact evaluation of a results-based financing (RBF) programme in Northern Uganda. The purpose of the programme was to improve health of disadvantaged post-conflict communities in support of private-not-for-profit providers. The team was unable to establish causality of the impact of RBF on health services or utilisation due to irreducible fundamental baseline differences between the project and the control groups.

We classified two out of nine reliable evaluations as not relevant. The evaluation of an empowerment programme in Uganda found positive impacts on intermediate outcomes. However, the sample used was scarcely representative. There were delays in the project implementation and in the evaluation, and the project underwent a number of changes during the evaluation. The project was unable to assess the impact of the

intervention on welfare outcomes and focussed its policy recommendations on the achievement of process outcomes only. The management response stated that by the time the recommendations became available, they were no longer relevant or had already been implemented by the project. The evaluation of another empowerment programme in Zambia found inconclusive evidence. There were delays in project implementation so that the planned impact evaluation became an 'interim' evaluation of the 'direction of travel' rather than the expected final assessment, thus informing a future study but not project implementation. The evaluation was unable to conclusively answer the five questions posed by the funder because the programme did not have the time to produce its impact.

Table 5: Reliability, Relevance and Policy Impact of DFID Evaluations

Study	Impact of intervention	credibility	relevance	Policy impact
Delivering Reproductive Health Results though non-state providers in Pakistan (DRHR)	No impact found	Yes	Yes	Yes
Evaluation of the Developing Operational Research Capacity in the Health Sector Project (ORCHSP)	Positive impact	No	No	No
Evaluation of the Pilot Project of Results-Based Aid in the Education Sector in Ethiopia (RBAESE)	No impact found	No	No	No
Social and Economic Impacts of Tuungane	No impact found	Yes	Yes	No
Independent Impact Assessment of the Chars Livelihoods Programme – Phase 1 (CHARS1)	Positive impact	No	No	No
Independent Impact Assessment of the Chars Livelihoods Programme – Phase 2 (CHARS2)	Positive impact	Yes	Yes	Yes
Evaluation of Uganda Social Assistance Grants for Empowerment (SAGE)	No Impact	Yes	No	No
Adolescent Girls Empowerment Programme, Zambia (AGEP)	Inconclusive	Yes	No	Not available
Impact evaluation of the DFID Programme to accelerate improved nutrition for the extreme poor in Bangladesh: Final Report (IEAINB)	No impact found	Yes	Yes	Yes
Investing in communities achieves results: findings from an evaluation of community responses to HIV and AIDS	mixed evidence of success	Yes	Yes	Not available
Independent Evaluation of the Security Sector Accountability and Police Reform Programme (SSAPR)	Impact on intermediate outcomes	Yes	Yes	Yes
Post Conflict Development in Northern Uganda - Results Based Financing to Private Not for Profit health facilities	Inconclusive	No	Yes	Partial (some learning about process and output differences)
Child Development Grant (CDG) Programme (Midline evaluation)	Impact on intermediate outcomes	Yes	Yes	Yes

Five evaluations had a significant policy impact. The evaluation of a health intervention through non-state providers in Pakistan delivered the 'surprising' result that the intervention had no impact on access to health care, which led to a re-examination of the whole intervention. The evaluation of a livelihood programme in Bangladesh encouraged the management to continue the intervention and to introduce changes to improve effectiveness. The 'sobering' results produced by the evaluation of a programme to accelerate improved nutrition for the extreme poor in Bangladesh led to a reconsideration of the DFID support strategy to nutrition interventions in Bangladesh. The evaluation of a policy reform programme in Zambia found a number of positive impacts on knowledge, attitudes and practices around policing that management accepted and committed to use in future programmes.

Two evaluations appeared at first not to have a policy impact: the Tuungane and the SAGE evaluations. Tuungane was a £90 million community development programme implemented by the International Rescue Committee (IRC) and CARE since 2007 in the DRC, based on community-driven reconstruction. The programme was split into 2 phases: Phase I ran from 2007-10 and phase II ran from 2010-14. The evaluation was unable to demonstrate an impact on behavioural change. According to the management response, the authors did not identify the reasons for the lack of impact and few policy recommendations were made.

The *Evaluation of Uganda Social Assistance Grants for Empowerment (SAGE)* tested two social transfers: the Vulnerable Family Support Grant (VFG), using a composite index to determine eligibility and transfers grants to households; and the Senior Citizens Grant (SCG), using age to determine eligibility and transfers grants to individuals. The evaluation showed significant positive changes in the lives of beneficiaries over two years using a combination of regression discontinuity design, propensity score matching, and difference-in-difference approaches. No management response was available for this study and we could not tell whether it was instrumental in changing policy, though the strength of the policy recommendations would suggest it was. However, in the absence of further information we coded the policy impact of this intervention as not available.

There are some limitations in using management responses to assess the policy impact of a study. First, it offers a partial view reflecting an exclusive attention to the implementation of specific programmes. Second, it represents policy impacts that are time bound to the immediate aftermath of a study or project. In other words, management responses are more appropriate to assess the instrumental use of evaluations then their conceptual use. As the case study analysis of the impact of the Tuungane evaluations shows (see Box 2), policy impacts of an evaluation may occur in the long term and inform decisions by other agencies than the original study commissioners. A deeper reading of the impact of the Tuungane evaluation, based on interviews with researchers and users, concluded that the study had a policy impact. As we will discuss in Section 7, 3ie employs a different measure of policy impacts, which considers both the instrumental and conceptual use of evaluations. For this reason, the policy impact assessments of 3ie and DFID studies are not comparable.

Box 2 The long term policy impact of the Tuungane evaluation

At first sight, the Tuungane evaluations appeared to be a classic case of negative deviance: a highly reliable and relevant study without a policy impact. The management response to the study reported that recommendations would not be used because the causes for the lack of impact of the intervention had not been explained. We decided to investigate the issue with the principal author of the study as well as with the original author of the management response and we gained a different understanding of the policy impact of the study.

The author of the evaluation claimed that the evaluation had been very influential, particularly at the World Bank, as the study results were consistent with those of evaluations of similar interventions carried out in other countries. The author of the management response explained the concerns raised in the management response in two ways. Firstly, it had been felt that the study may have benefited from a complementary qualitative evaluation of the findings to understand the reasons for a lack of impact. Secondly, the conclusions of the evaluations were considered premature given the nature and complexity of the intervention the effects of which were expected to be realised in the long term. These concerns notwithstanding, the author of the management response confirmed that 'the evaluation was hugely influential in subsequent funding investments made by DFID in other community-driven developmentirc initiatives.' The conversation also highlighted that there are other documents such as DFID Annual Reviews, Business cases, and Programme Completion Reviews that represent more comprehensive and instrumental reference documents to assess policy impact. One simple lesson from this case study is therefore that policy impact of an evaluation is hard to detect and that it requires a long investigation of relevant documents and interviews with main stakeholders.

It is also interesting that only one out of nine reliable studies found a positive impact. Two studies found a positive impact but the evidence provided was not reliable. Three studies found a positive impact but only on intermediate outcomes. All the remaining studies found no impact of the interventions on final outcomes. It is also interesting that all studies were assessing impacts on a plurality of outcomes and sometimes for several sub-groups in the population. We found no attempt to cherry-pick the results or of selective reporting, but evaluators and management often struggled to make sense of a large number of effects observed on several dimensions and the results were difficult to interpret in a coherent way.

Characteristics associated with Success in DFID Evaluations

Five evaluations were pilot interventions. The remaining evaluations assessed existing programmes with the intent of improving or scaling-up the interventions or for accountability purposes. Nine evaluations included a good theory of change, but only three quoted systematic reviews of similar interventions.

Success at the planning stage depends on factors such as: the skill set of the evaluators, transparency of the evaluation, stakeholder engagement, and the implementation of pilot or formative studies before starting the full evaluations. None of the studies reported an evaluability assessment, and none of the newly implemented projects had a pilot or a formative evaluation, and nothing is known about engagement with stakeholders. It is impossible to say to what extent these activities did not take place or were simply not reported. We know however that all evaluations funded by DFID are required to go through an independent quality assurance service for review.

Eight programmes faced difficulties at the implementation stage which resulted in delays. Findings from the RBA pilot were not communicated in a timely manner to the regions to appreciably affect students' performance. In the Tungane evaluation, political tensions led to the loss of one province and some regions were inaccessible for safety reasons. In the CHARS evaluation, the initial implementation of the monitoring system was rudimentary, and cohorts were mixed within villages. In the adolescent empowerment programme in Zambia, the voucher scheme took almost two years to agree with the government. In the integrated nutrition programme in Bangladesh, procurement and distribution of micronutrients were delayed.

The Security Sector Accountability and Police Reform Programme did not state its short- and long-term goals until quite late in the implementation period, and thus there was little time to work collaboratively. The 'Post Conflict Development in Northern Uganda – testing Results Based Financing to Private Not for Profit health facilities' intervention was affected by a limited comparability between the treatment and comparison groups, which ultimately led to a change in the Theory of Change and a reduction in the budget.

The Child Development Grant (CDG) was affected by security issues in the intervention area, which prevented the team from visiting several communities, with the resulting imperfect coverage of eligible women. This study and the Bangladeshi nutrition programme, were also affected by a major shock. Attrition above 20% of the baseline sample was reported as a problem in three studies. No study reported data on take-up. Sample sizes were in the order of thousands of observations in most cases but two evaluations collected information from fewer than 300 observations. We were unable to tell whether the latter studies were underpowered from the available data, but evaluations of social interventions employing small samples are more likely to find false negative results. There is also a risk that small studies produce false positive results through a process of reporting bias (Gelman and Weakliem, 2009). Small studies of a given intervention will produce a large variety of results around the true project effect (more and less positive), but often times it is only the more positive results that end up being reported and published.

All studies, with three exceptions, were mixed method evaluations that included a qualitative component in the form of interviews, focus group discussions, process evaluations or contribution analysis. All studies, with three exceptions, conducted some sub-group analysis mostly by gender and age of project participants. Only three studies, namely the Bangladeshi nutrition study, the Child Development Grant and the Evaluation of the Security Sector Accountability and Police Reform Programme, analysed intermediate outcomes of the interventions.

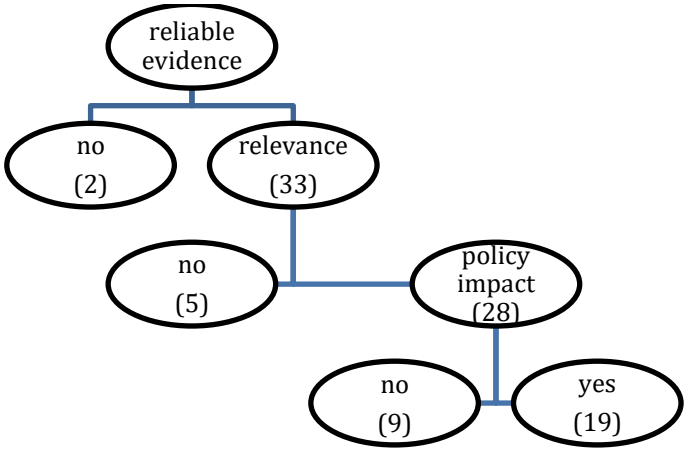
Section 5

3ie Impact Evaluations

Successful 3ie Studies

Evaluations funded by 3ie are selected among a large number of proposals based on their policy relevance among other criteria. In addition, the evaluations are monitored from design to publication in order to avoid errors in planning, implementation and analysis. As a result, credibility and relevance of 3ie-funded evaluations tend to be high. [9] Credibility and relevance were assessed by one of the authors based on extensive notes from 3ie reviewers. 3ie routinely conducts its own assessment of the policy impact of the evaluations it supports through interviews with researchers and policy-makers and classifies studies as used instrumentally ('changing policy or programme design' and 'taking a successful programme to scale'), conceptually ('informing discussions of policies and programmes' and 'improving the culture of evaluation evidence use and strengthen the enabling environment'), or as not being used. In what follows we used the 3ie classification to assess the policy impact of 3ie-funded studies. Unfortunately, the information available on DFID studies did not allow to perform the same classification for the evaluations funded by 3ie and, as already observed in the methodology section and again in Section 6, the policy impact of 3ie and DFID studies are not comparable.

Figure 7: Reliability, Relevance and Policy Impact (3ie Evaluations)



Only two evaluations, out of the 35 selected, produced non-reliable evidence (Figure 7). The first was an evaluation of the impact of a micro-irrigation innovation on household income, health and nutrition. The study had enormous limitations relating to the comparability of three cohorts, and attrition was not properly addressed. The second was a study assessing the effect of network structure on the diffusion of health

9 The full set of selection criteria used by 3ie includes 3ie section criteria also include quantitative and qualitative evaluation skill set of the study team, their sector expertise, grant management experience, policy relevance of the proposed study and the value for money.

information. The study was designed in such a way that the desired estimation of the causal impact was impossible.

We found five non-relevant evaluations. A study of an intervention promoting savings among poor households in Sri Lanka was unable to provide policy recommendations because it was affected by low take-up and delays.[10] A study of an intervention aiming at increasing long-term savings with a pension product was affected by implementation problems; it did not produce the expected results, and was of no policy use. A study of an intervention promoting the use of cook-stoves to prevent indoor pollution in Northern Ghana faced so many implementation problems that the results were few and difficult to interpret. A study promoting male circumcision in Malawi produced very limited results because of very low take-up among the target population. Similarly, a study of an intervention offering weather index-based insurance to poor farmers in two Indian states failed because few farmers were interested in the intervention.

We found implementation issues and low take-up to be the factors more frequently associated with poor relevance. The cook-stove and the mobile banking evaluations were affected by factors outside the control of the researchers or by factors not well understood at the design stage. The designs of these studies could potentially have been improved through a better engagement with relevant stakeholders and with an evaluability exercise to assess potential risks. Other studies were affected by implementation delays and low take-up. Finally, some interventions were testing totally new interventions which ended up being ineffective. Some innovation studies are designed in such a way to be scarcely relevant when successful, and poorly informative when unsuccessful (see Box 3).

Table 6 Credibility, relevance, and policy impact of 3ie-funded studies

Study	Impact of intervention	credibility	relevance	Policy impact
Age at marriage, women's education and mother and child outcomes in Bangladesh	No impact	yes	yes	yes
Shelter from the storm Upgrading housing infrastructure in Latin American slums	Positive impact	yes	yes	yes
Chlorine Dispensers in Kenya: Scaling for Results	Mixed results	yes	yes	no
Estimating the Impact and Cost-Effectiveness of Expanding Secondary Education in Ghana	Positive impact	yes	yes	no
The impact of daycare on maternal labour supply and child development in Mexico	Mixed results	yes	yes	yes
Monitoring and Assessing the Impacts of KickStarts Low Cost Farm Equipment on Poverty Reduction in Africa	No impact	no	no	no
Improved cook-stoves in the Tumu region of Ghana	No impact	yes	no	no
Learning and Growing in the Shadow of HIV/AIDS: A Prospective Randomized Evaluation of the Effects of Escolinhas on Young Children in Mozambique	Positive impact	yes	yes	yes
Environmental and Socioeconomic Impacts of Mexico's Payments for Ecosystem Services Program	Mixed results	yes	yes	yes

[10] Shortly after the demonstrations began, a glitch was discovered in the IS platform on which the product was built. The mobile operator's account was debited each time a deposit was made and the bank suspended use of the system. There were delays in the signing of MOUs between banks and mobile services. Finally, a credit card fraud made the bank more cautious about the access given to the software company which caused some further delays in the launch of the product

Courting Safe Behaviors: Testing Courtyard-based Safe Water and Hygiene Interventions in Urban Bangladesh	Mixed results	yes	yes	no
Wage South Africa	Mixed results	yes	yes	yes
Assessing the effectiveness of improved sanitation on diarrhoea, nutritional status and helminth infection: a cluster-randomized, controlled field trial in Orissa, India	No impact	yes	yes	no
An impact evaluation of information disclosure on elected representatives performance: Evidence from rural and urban India	Positive impact	yes	yes	yes
Smallholder access to weather securities: demand and impact on consumption and p	No impact	yes	no	no
A Randomized Evaluation of the Effects of An Agricultural Insurance Program on Rural China	Mixed results	yes	yes	yes
The Impact of Mother Literacy and Participation Programs on Child Learning	Positive impact	yes	yes	yes
The Economics and Psychology of Long-term Savings and Pensions: A Randomized Experiment Among Low-income Entrepreneurs in Maharashtra, India	Mixed results	yes	no	no
Scaling up Male Circumcision Service Provision	No impact	yes	no	yes
The Diffusion of Health Knowledge through Social Networks: An Impact Evaluation	Mixed results	no	no	no
Property Tax Experiment in Punjab, Pakistan	Positive impact	yes	yes	yes
Enabling Micro-savings Through Bank-Linked Mobile Phones and Mobile Banking in Sri Lanka	No impact	yes	no	no
Micro Entrepreneurship Support Program in Chile: Impact Evaluation	Positive impact	yes	yes	no
Targeting the Ultra Poor: an impact evaluation of the BRACs Graduation Model in Ghana	Positive impact	yes	yes	yes
Impact of free availability of public childcare on labour supply and child development in Brazil	Positive impact	yes	yes	no
Aid and accountability: governance effects of a community-driven reconstruction program in eastern Congo	No impact	yes	yes	yes
Improving Targeting in Conditional Cash Transfer Programs: A Randomized Evaluation of Targeting Methods in Indonesia's CCT Program	Positive impact	yes	yes	yes
Livelihood empowerment against poverty program impact evaluation	Mixed results	yes	yes	yes
The impact of Tanzania's Joint Forest Management programme on livelihoods, governance and forests	Mixed results	yes	yes	no
Business training services	Positive impact	yes	yes	yes
Evaluation of Centres for Infant Development: an early years intervention in Colombia.	No impact	yes	yes	yes
Evaluating the impact of closing a community managed schools programme in El Salvador	No impact	yes	yes	no
Impact evaluation of the Living Side-by-Side, peacebuilding programme in Kyrgyzstan	Positive impact	yes	yes	no
Economic Growth And Risk Reduction in Malawi's Social Cash Transfer Scheme	Positive impact	yes	yes	yes
The impact of the Productive Safety Net Programmme on the schooling and nutrition of children	No impact	yes	yes	yes
Impact evaluation of the programme supporting development of Menabe and Melaky regions in Madagascar	Positive impact	yes	yes	no

Of the 28 studies identified as relevant and reliable, only six were influential in 'changing policy or programme design', and one was influential in 'taking a successful programme to scale.' Seven studies 'informed discussions of policies and programmes', 4 studies improved 'the culture of evaluation evidence use and strengthen the enabling environment', while the remaining 9 studies appeared to have no policy impact. 3ie

evaluations were rarely directly commissioned by a funder to answer a precise question. 3ie was established precisely with the goal of supporting impact evaluations that would have not otherwise been funded. It should be no surprise therefore that some 3ie evaluations were not designed to influence decisions directly.

Box 3: Publication ≠ Relevance

The publication of an evaluation in a peer-reviewed journal is sometimes used an indicator of the quality of the study. However, published studies are credible and interesting but not necessarily relevant. Some impact evaluations are designed to test research ideas. But where do ideas come from? In the words of Karlan and Appel of *Innovations for Poverty Action* (Karlan & Appel, 2016a), ideas are 'perhaps extensions or continuations of an existing theory, or inspired by the results of other research or from experiences in neighbouring countries or countries far away.' Ideas often come in the form of hypotheses about incentives influencing behaviours (the 'nudge'). For example, in a famous study mothers were offered two pounds of lentils and a set of stainless steel plates if they vaccinated their children. The incentive had a large impact on a full cycle of vaccination rates (a sevenfold increase to 38%, though it had no impact on the first three injections), and the authors (Banerjee & Duflo, 2011) judged this programme 'to be the most effective we ever evaluated, and probably the one that saved the most lives.' One criticism of these type of studies is that they produce knowledge which is too narrow and local to be useful for policy (Deaton, 2010). If a similar study is successful, it shows that a particular incentive has a specific effect in a given population. But what do we know about the impact of other types of incentives and what do we know about the impact of the same incentives in other populations? On the other hand, if the study is unsuccessful it provides no other recommendation apart from confirming that a particular type of incentive does not work. It has been suggested that these evaluations are 'mechanism experiments' testing theory having a general validity (Bates and Glennerster, 2017). Mechanism experiments are evaluations testing a particular link in the causal chain of an intervention rather than its effectiveness (Ludwig et al, 2011). They are designed not to evaluate the effectiveness of any specific intervention but to uncover the operation of behavioural mechanisms that underpin the theory behind an intervention. If the theory is a mid-level theory that is sufficiently general to abstract from the specific circumstances of the intervention but without taking the form of an all-explaining general theory, then mechanism experiments may provide policy recommendations that are externally valid (Davey et al, 2018). These type of experiments however are rare and many behavioural change evaluations are effectiveness trials testing new incentives. Journals are keen to publish this type of study because they are innovative and sometimes defy existing theories, but the policy relevance of these studies is questionable. 3ie funded a number of evaluations that were subsequently published in peer-reviewed journals that were considered of little relevance by 3ie internal and external reviewers. Reviewers of academic papers and reviewers of impact evaluations use different standards and policy relevance does not appear to figure highly among the criteria used by academic reviewers.

One difficulty in the use of the 3ie measure of policy impact is that it is not clear to what extent the reported lack of impact is a reflection of evidence of lack of impact or rather absence of evidence of impact. The latter is more likely because evidence of impact is difficult to collect through interviews and, as discussed in the case of DFID evaluations, policy impacts can occur at different times and places. The amount of information on policy impact that can be collected after the publication of an evaluation study is limited and partial by construction. While 3ie's efforts to document impacts is more in depth than DFID, even for 3ie more investigative research through documents reviews and interviews would be needed in order to identify all impacts from evaluations. Even if evidence of lack of policy impact is obtained, it cannot be excluded that a study will become influential in the future or that was influential in areas that we were not able to observe.

Less than half (14) of the 35 studies considered, found a clear positive impact of the intervention. Eleven studies found no impact at all, while nine studies found 'some' impact on some of the outcomes observed. Many studies found mixed results because they tested several outcomes at the same time. In the absence of a pre-analysis plan setting out the hypotheses to be tested, it is difficult to interpret the results of these evaluations. Lacking a pre-specified set of hypotheses to test, the results can be reported selectively by the researchers in a cherry-picking exercise led by pursuit of statistical significance. There is evidence, at least in economics, that some researchers may refine their analyses until reaching statistical significance in order to increase the probability of being published (Brodeur et al., 2016). A similar behaviour by researchers conducting impact evaluations is to be expected. Of the nine studies reporting 'some' impact, only one had a pre-analysis plan and the results produced by these studies were difficult to interpret.

Characteristics of successful 3ie-funded evaluations

Table 7 illustrates the characteristics of 3ie evaluations. With the exception of team's skills, which is an index of team's qualification running from 0 to 4, and of attrition and take-up, which are averages of percentages in each study, all characteristics considered are binary. They were coded as zero or one for each evaluation and we reported in the table the average, which can be interpreted as a percentage rate. As noted in Section 3, we were not able to code the studies funded by DFID with the same level of detail because of lack of data and therefore we were not able to build a similar table for DFID-funded evaluations.

We first discuss briefly the average characteristics of 3ie evaluation before comparing the differences in characteristics of successful and unsuccessful evaluations. Only a quarter of evaluation designs included a well-defined theory of change and only about a third quoted a systematic review. Good team skills are a 3ie funding criterion and rated on a scale from 1 to 4. Unsurprisingly, team's skills ratings were quite high among funded evaluations. Only two studies had an evaluability assessment. Nearly half of the studies were preceded by a pilot or formative study assessing the feasibility of the intervention.

Table 7: Characteristics associated with Success in 3ie Studies

Determinant	Successful	Unsuccessful	All
Design and planning			
Well-designed theory of change	0.17	0.29	0.23
Systematic review quoted	0.17	0.53	0.34
Team's skills	3.17	2.65	2.91
Evaluability assessment	0.00	0.12	0.06
Pilot or formative study	0.39	0.53	0.46
Implementation			
Implementation issues	0.33	0.47	0.40
Implementation delays	0.33	0.53	0.43
Letter of variation	0.94	0.94	0.94
Changes to design	0.00	0.24	0.11
Take-up (average)	0.87	0.61	0.74
Attrition (average)	0.05	0.17	0.12
Sample size	0.89	0.47	0.69
IRB approval	0.61	0.53	0.57
Analysis			
Mixed methods	0.83	0.88	0.86
Subgroup analysis	0.72	0.71	0.71
Intermediate outcomes	0.83	0.82	0.83
Dissemination			
Policy impact plan	0.83	0.94	0.89
Policy brief	0.17	0.12	0.14
Presentations	0.50	0.41	0.46
Media	0.28	0.24	0.26
Publications	0.83	0.88	0.86
Budget	2.00	2.00	2.00
Cross-cutting factors			
Shocks	0.11	0.12	0.11
Study registration	0.11	0.24	0.17
Pre-analysis plan	0.17	0.12	0.14
Implementer's endorsement	1.00	1.00	1.00
Stakeholders meetings	0.50	0.65	0.57
Stakeholders trainings	0.17	0.12	0.14
Peer-review group	0.78	0.65	0.71
Advisory panel	0.72	0.82	0.77

Note: all variables are binary with the exception of team's skills, which is measured on a scale from 0 to 4. Sample size is considered sufficient if larger than 1,000 observations. Take-up and attrition are averages of study-specific rates.

Piloting, however, did not prevent implementation problems and delays, which occurred in 47% and 53% of cases respectively. That delays are a common issue is evidenced by the number of letters of variations issued (94% of cases), though only about 10% of studies changed their original design. The average attrition rate (12%) is not too high and 80% of studies had an attrition rate below 20%, but one third of the studies did not report attrition rates. Only 10% of evaluations were affected by major

shocks. About 70% of evaluation had a reasonable sample size. Surprisingly, only slightly above 50% of the studies were approved by an institutional review board.

The proportion of registered studies and of studies with a pre-analysis plan was very small (less than 20%). More than three quarters of evaluations had a peer-review group or an advisory panel. 3ie promotes the engagement of research teams with project managers and policy makers, so that implementers' endorsement is universal and the proportion of evaluation teams conducting meetings with the implementers was very high.

At the analysis stage, the use of sub-group analysis, the analysis of intermediate impacts and the use of mixed methods approaches was very common. All evaluations had a policy influence plan, though this did not always translate in dissemination as shown by the relatively low proportion of studies presented at public events, disseminated through policy briefs or in the media. On the other hand, nearly all evaluations funded by 3ie were published in peer-reviewed journals.

Our sample was too small to perform statistical tests of differences in characteristics between successful and unsuccessful studies. However, we could compare the average levels of characteristics in the two groups to see if there are any glaring differences. First, differences were not large, which implies there was no obvious factor strongly associated with our definition of success. In our sample of 3ie studies, the use of theories of change, systematic reviews, evaluability assessments and pilots was not associated with the success of the study. The only characteristic that was more prominent among successful studies at the design stage was team's skills.

Some interesting differences appear at the implementation stage. Successful evaluations in this sample were less affected by implementation issues and delays. They had lower attrition rates and larger sample size. Unsuccessful evaluations had lower take up rates and more frequent changes in evaluation design. Characteristics at the analysis and dissemination stage were nearly identical in the two groups. Similarly, other cross-cutting characteristics such as stakeholder engagement (endorsement, meetings and training), transparency (registration and pre-analysis plan), quality assurance mechanisms (advisory panels and peer-review groups) and shocks are hardly different in the two groups.

We also employed a truth table analysis to see whether any factor was sufficient to success (being present in some cases when the study was successful but never when it was not successful) or necessary (being present in all successful case but also in some unsuccessful ones). This analysis however did not add further insights. Only two factors were sufficient, but not necessary for *lack* of success: evaluability assessment and changes in design. These characteristics were present in some unsuccessful evaluations but absent in all successful ones.

Unsuccessful Studies

In this section we discuss cases in which in the course of an evaluation it became obvious that the study was likely to be unsuccessful. This situation is particularly challenging for both researchers and managers because a decision needs to be made

on whether the study should be discontinued or not. We only considered 3ie evaluations because we did not have sufficient information regarding on similar DFID cases.

Over the last ten years, 3ie discontinued 8 studies, out of 222 closed and ongoing evaluations. These evaluations were discontinued for various reasons. For example, a trial of an intervention to promote empowerment in Mauritania was discontinued at an early stage because of security concerns due to a sudden break out of hostilities. In another example, an evaluation of a debt relief mechanism in Andhra Pradesh was discontinued because of an inability to hire a survey firm. In addition, when the team went to the field to refine the survey instruments it became apparent that all villages in the state were being reached by the intervention thus preventing the establishment of a control group. In another example, the study was terminated at the analysis stage. An evaluation of an intervention promoting the diffusion of improved seeds in Uganda was commissioned to a team with limited experience of impact evaluations. Concerns about the design were raised early on by 3ie reviewers and the study started as an RCT, soon to became a difference-in-difference study and finally turned into a PSM study without ever adopting a convincing evaluation approach.

While knowing the causes of the failure of studies, many of which are documented by Karlan and Appel based on the experience of IPA (2016a), is useful, it is also interesting considering studies that could have been discontinued but ultimately were not. Mckenzie observes that implementation failures tend to snowball quickly and that researchers find it hard to know when to walk away (2016). Similarly, Karlan and Appel observe that 'of all the intellectual challenges that arise in the course of designing and implementing a rigorous research study, the greatest may be deciding when to pull the plug.' The decision of discontinuing a study is as difficult for a manager as it is for a researcher. Here, we summarise some of the arguments that have been put forward by 3ie reviewers and managers when deciding whether discontinuing a study or not. We hope that a discussion of these arguments might help avoiding common mistakes and inform future decisions.

One reason for not discontinuing a study was the sunk costs fallacy. It is difficult to discontinue a study in which many resources and efforts have been invested, sometimes over a number of years. Sunk cost is a standard fallacy of logical reasoning: disbursements made in the past should not affect decisions about the future. In principle, managers and researchers should not allow past efforts to influence future decisions, which should only be based on an analysis of future benefits and costs. However, funders sometimes hope that circumstances will change or that researchers will devise alternative evaluation approaches. The sunk cost fallacy is difficult to resist.

A second common reason partly follows from the first. It was hoped that researchers may devise a change in design that would allow the completion of the study despite the change in environment and other difficulties. Many projects in international development end differently from what had been planned, and some degree of flexibility to adjust to changes in circumstance should be expected. For example, in a hypothetical case, a randomized trial may no longer be possible, a quasi-experimental

study or a qualitative study could turn out to be useful and researchers and managers could agree on a way forward.

A third argument is reputational risk. The success of an evaluation study is a responsibility of researchers as well as of their managers. Bringing a study to completion and meeting the targets despite all odds can avoid bad publicity or negative feedback. This is not a logical argument in favour of continuing evaluations of course but something to consider by managers of evaluations.

The publication of an unsuccessful study in an academic journal is another argument. Evaluations that fail at the implementation stage or that are not policy relevant (see Box 2) can still be published in academic journals. Academic researchers are highly motivated to publish studies as this is often tied to their job tenure. Managers and funders may sympathise with researchers and share their same motivation. An evaluation may not be discontinued to allow a journal publication even if it is known that its results are not policy relevant or even reliable.

Finally, some argue that a study should never be discontinued and should be carried out according to plan to the extent this is possible regardless of circumstances. The reason is that a discontinued study is forever forgotten and will not appear in reviews of evaluations. It is argued that the loss of the results obtained by a failed study will result in biased assessment of similar types of interventions. In addition, it is argued that much could be learnt from failures by other researchers if the study and its development were properly documented, if anything to prevent a similar failure to occur in the future.

We believe that while discontinuation should not be taken lightly it should be consistently considered. We do not believe that evaluations should be discontinued when they are not going to plan. Many evaluations take place in highly volatile and difficult contexts. A strict application of this rule would lead to the completion only of projects implemented and evaluated in simple environments. However, we also do not believe in continuing evaluations regardless of emerging circumstances simply for the purpose of documenting and avoiding reporting bias in research. A middle ground needs to be found between these two extremes and decisions should be taken on a case-by-case basis, though some general guidelines could be useful.

First, prospective publications in academic journals and reputational risk for the funding institution are not valid arguments in favour of continuing a study which is not set to provide reliable and relevant evidence. Academic journals publish studies that are not necessarily reliable and relevant and the reputation of funding institutions is better served by holding firmly to principles of rigour and relevance. Second, before discontinuing a study affected by low take-up, researchers and managers should consider the opportunity of applying different evaluation designs (including qualitative approaches), assessing impact on beneficiaries only (the 'treatment on the treated') rather than on the overall population, and explaining the reasons for low participation rates. Third, funders should establish monitoring processes to assess uptake and make informed decisions around implementation rollout or low take-up concerns. Funders should have designated 'go-no go' decision periods after some points in the evaluation.

These decisions would include a formal review of the study and a decision on the viability of the impact evaluation.

Section 8

Conclusions

We summarise here the main lessons of our review. Our analysis has a speculative nature and we make no claims of causality running from the factors considered to the success of the evaluations.

The overall success rate of impact evaluations was not very high. We worked with very small samples, and percentages can be misleading, but we found less than 50% of evaluations to be 'successful' by our definition. Meeting the triple goal of achieving credibility, relevance and policy impact is not a simple task. Expectations about success of impact evaluations should be set accordingly.

DFID studies tend to have higher policy impact while 3ie studies tend to score more highly on credibility. This is likely the reflection of the different goals adopted by the two institutions. 3ie is more inclined to fund evaluations informing global debates and increasing the stock of knowledge ('public goods'), while DFID is more inclined to commission evaluations with a view to directly improving programme implementation.

Our measurement of evaluation use is imprecise. Even 3ie's efforts to document evidence use through interviews with researchers and implementers are insufficient. The DFID case studies of reliable evaluations with no policy impact, have shown that evaluations can be extremely influential but not always at the expected time and place. The lack of evidence of policy impact does not imply the absence of policy impact, because this can only be detected through detailed case studies. This points to a fundamental problem for any attempt to measuring policy impact using simple metrics or quick data collection methods by interviews and documents reviews.

3ie has funded many 'innovations' - evaluations of interventions never experimented before. These evaluations were often preceded by pilots in order to avoid failures resulting from low acceptability and feasibility. However, there is always an element of risk in 'innovations' and not everything can be planned in advance or piloted. The experience of 3ie also shows that 'innovations' evaluations are not always policy relevant.

At the planning and the design stage, the composition of the team of evaluators and their skills appear to be an important factor associated with success, while other factors often stressed in the literature, such as stakeholders' engagement, a solid theory of change, and peer-review groups, appear to be less relevant. We note that, as already observed, our study is purely descriptive and based on a very small sample and these statistics are not unbiased and generalizable.

The low rate of registration of evaluations and of pre-analysis plan does not seem to affect the success of the evaluations but does have a negative effect on policy

recommendations. Many evaluations report 'some' impacts after testing multiple hypotheses for several outcomes in such a way that the results of the interventions are difficult to interpret. Registration and pre-analysis plan would help researchers to clarify the goals of their study and to set more clearly the standards used to assess impact.

Most problems affecting impact evaluations seem to emerge at the implementation stage. We found many evaluations affected by delays in programme approval, project modification, implementation delays, lack of financial and/or human capital to roll out the programme as planned, lack of buy-in from lower level and last-mile workers, service delivery glitches, external shocks and low take-up. Low take up in particular is one of the main sources of unsuccessful evaluations with implications on the credibility of the estimation of project effects and on their external validity.

In our sample of 3ie studies, successful evaluations had overwhelmingly obtained their data from large samples. Sample size is a crude measure of the quality of a study but we observe that evaluations relying on small samples (less than 1,000 project observations) are more likely incur into problems. Small sample sizes are more likely to produce null or inconclusive results because of low statistical power, but they can also reflect poor planning and design or a more general lack of skills and experience in the evaluation team. Recently it has been suggested that the p-value for statistical significance should be reduced from 5% to 0.5% for new discoveries with the goal of reducing the publication of false positives (Benjamin et al, 2017). Since statistical significance is a function of sample size, the suggestion is equivalent to asking that evaluation studies should be designed using large samples. Statistical significance cannot be the only criterion to assess the validity of new evidence (McShane et al, 2017) but funding agencies may become more demanding in terms of sample sizes.

The use of mixed methods, sub-group analysis and the analysis of intermediate impact does not appear to make studies more successful or policy relevant. On the contrary it was observed that in the absence of registration and pre-analysis plan the production of many results can make the interpretation of impact more difficult.

Whether an evaluation should be discontinued when facing implementation problems is a decision that should be taken on a case-by-case basis. We do not recommend that evaluations should be discontinued when sticking to plans is impossible, or that they should be led to completion at all costs regardless of the problems faced. Managers and researchers should avoid considerations of reputational risk, publication of research or sunk costs when making decisions regarding discontinuations, and a process should be established to consider the opportunity for discontinuing evaluations.

Section 9

Relevance for CEDIL

This paper can inform new strands of work by CEDIL in three areas: guidelines for selecting and quality-assuring impact evaluations, protocols for managing and adapting evaluations in difficult contexts, and indicators for tracking policy impact. The proposed

work cuts across the CEDIL themes and could be developed by the Research Directorate, by the Programme Directorate, by some of the consortium members, or by a combination of them.

The first proposed area of work is the development of guidelines for selecting and monitoring impact evaluations funded by CEDIL. The guidelines would be directed at both managers and researchers and would include selection criteria for evaluation proposals as well as protocols for monitoring and quality-assure evaluation studies. Similar guidelines already exist and are currently in use by 3ie and DFID. Our review could help updating and refining elements of these guidelines. For example, our review has identified the following factors associated with success:

- Impact evaluations should be closely monitored and quality-assured from the design stage to the publication and dissemination of results
- Evaluation proposals should be scrutinised to ensure their relevance and to avoid 'researcher's capture', whereby studies are conducted to answer some 'new' and intellectually challenging, but practically irrelevant, questions
- All evaluations should be registered and have a pre-analysis plan
- Evaluation questions should be based on a solid theory and deep understanding based on the available evidence
- Evaluation teams should be multidisciplinary and adequately skilled
- Evaluations should rely on large samples of observations. Asking teams to conduct power calculations is not sufficient. Power calculations should be conducted using margins of errors and expected effect sizes more conservative than those normally in use. There is ample evidence from epidemiological studies that studies are often underpowered even when power calculations have been conducted (Turner et al, 2013)

A second potential area of work relates to the complexity of the research themes and contexts in which CEDIL intends to operate. CEDIL is committed to work in neglected areas, such as fragile states. Impact evaluations face implementation difficulties and delays even in simple environments. The difficulties to successfully complete impact evaluations funded by CEDIL are bound to be higher. Our review has shown that changes to impact evaluation designs are the norm rather than the exception. The organisational experience summarised in our review suggests that strictly adhering to design protocols is often difficult if not impossible. If we were to strictly adhere to original evaluation plans, only few evaluations would be completed, particularly in difficult environments. Researchers have to be prepared to adapt their designs to emerging circumstances, while managers have to be prepared to make decisions about modifying and discontinuing evaluation studies. We propose therefore that guidelines should be developed for performing some basic risk analysis at the design stage of an evaluation and that research teams should be encouraged and supported to design contingency plans including alternative study designs as circumstances are changing. We are also proposing the development of guidelines and protocols to support researchers and managers to decide when it is the right time to modify or discontinuing a particular study.

A final potential area of work is in assessing the policy impact of evaluations. Our review found that just because an impact evaluations is reliable and relevant does not mean they always have impact on policy. We showed that the search for policy impact can be elusive. As impacts can be subtle or occur in the long term, the search for impact can be very demanding and never-ending. We propose therefore that CEDIL should promote instead an unpacking of the 'loose' theory behind the policy impact of evaluations. This work would require firstly reviewing the existing knowledge and secondly to theorise and conceptualise policy impact. Rather than striving to observe an impact difficult to observe, we propose to identify and test specific links of the policy impact theory of change and employing intermediate indicators to monitor policy influence.

References

Alkin, M.C., and King, J.A. (2017). Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use, American Journal of Evaluation, 38(3), 434-50.

Banerjee, A., & Duflo, E. (2011). *Poor Economics: A radical rethinking of the way to fight global poverty*. New York: PublicAffairs.

Bates, M. A, and R. Glennerster (2017) The Generalizability Puzzle, Stanford Social Innovation Review, Summer 2017, https://ssir.org/articles/entry/the_generalizability_puzzle.

Benjamin D.J. et al (2018) Redefine statistical significance, *Nature Human Behaviour*, 2, 6-10.

Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2016). Declaring and Diagnosing Research Designs. *unpublished manuscript*.

Briceno, B., Cuesta, L., & Attanasio, O. (2011). Behind the scenes: managing and conducting large scale impact evaluations in Colombia. Journal of Development Effectiveness, 3(4), 470-501.

Brodeur, A. *et al.* (2016) 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics*, 8(1), pp. 1–32. doi: 10.1257/app.20150044.

Brown, J., Cartridge, R., Davies, S., ul Hassan, E., Hsia, C., Lal, T., Mackison, D., Menon, R., Moratti, M., Murunga, V., Le Nestour, A., Whitehead, L., and H. White (2018), Research Uptake in Policy-Making: from Papers to Policy, CEDIL Inception paper no. 14: London

Campos, F., Coville, A., Fernandes, A. M., Goldstein, M., & McKenzie, D. (2012). *Learning from the Experiments That Never Happened: Lessons from Trying to Conduct Randomized Evaluations of Matching Grant Programs in Africa*. Washington DC: The World Bank.

Cousins, J.B. & Leithwood, K.A. (1986) Current Empirical Research on Evaluation Utilisation. *Review of Education Research*, 56(3), 331-364.

Culbertson, M.J. McCole, D., McNamara, P.E. (2014). Practical challenges and strategies for randomised control trials in agricultural extension and other development programmes, Journal of Development Effectiveness, 6(3), 284-299.

Cuong, N. V. (2014). *Impact evaluations of development programmes: experiences from Viet Nam. 3ie Working Paper 21*. New Delhi: International Initiative for Impact Evaluation (3ie).

Davey, C., Hassan, S., Cartwright, N., Humphreys, M., Masset, E., Prost, A., Gough, D., Oliver, S., Bonell, C. and J. Hargreaves (2018) Designing evaluations to provide evidence to inform action in new settings, CEDIL Inception paper no. 2: London

Davies, R. (2015). Evalauability assessment. Retrieved from http://www.betterevaluation.org/en/themes/evaluability_assessment

Davies, R. (2016). Evaluating the Impact of Flexible Development Interventions Using a 'loose' Theory of Change: Reflections on the Australia-Mekiong Ngo Engagement Platform. ODI Methods Lab.

Davies, R. (2018). Representing Theories of Change: Technical Challenges and Evaluation Consequences, CEDIL Inception paper no. 16: London.

de Silva, S.J., Paci, P., and Posadas, J. (2015) A checklist to avoid pilot failures: lessons from a set of women's economic empowerment initiatives. Journal of Development Effectiveness, 7(1), 90-110.

Duflo, E. (2017). The Economist as a Plumber.

Epstein, D., and Klerman, J.A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. Evaluation Review, 36, 375-401.

Ferguson, M. (2017). Turning lemons into lemonade, and then drinking it: Rigorous evaluation under challenging conditions. Retrieved from https://researchforevidence.fhi360.org/turning-lemons-lemonade-drinking-rigorous-evaluation-challenging-conditions

Fleischer, D.N. & Christie, C.A. (2009) Evaluation Use: Results from a Survey of U.S. American Evaluation Association Members. American Journal of Evaluation, 30(2), 158-175.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. American Scientist, 97: 310-316.

Gelman, A. and E. Loken (2013). The Garden of Forking Paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time, Technical report, Department of Statistics, Columbia University, New York.

Glennerster, R. (2017). The Practicalities of Running Randomised Evaluations: Partnerships, Measurement, Ethics, and Transparency. In E. Duflo & A. Banerjee (Eds.), *Handbook of Economic Field Experiments* (Vol. Vol. 1, pp. 175-243): North Holland.

Glennerster, R., & Takavarasha, K. (2013). *Running Randomised Evaluations: A Practical Guide*. Princeton, New Jersey: Princeton University Press.

Gough, D. (2007) Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. In J. Furlong, A. Oancea (Eds.) Applied and Practice based Research. Special Edition of Research Papers in Education, 22, (2), 213-228

Greenberg, D. Barnow, B.S. (2014) Flaws in Evaluations of Social Programs: Illustrations From Randomised Controlled Trials, Evaluation Review, 38(5), 359-387.

Gueron, J. M., & Rolston, H. (2013). *Fighting for Reliable Evidence*. New York: Russel Sage Foundation.

Hearn, s. and A.L. Buffardi (2016) What is impact? A Methods Lab publication. London: Overseas Development Institute.

Herbert, J.L (2014) Researching Evaluation Influence: A Review of the Literature. *Evaluation Review*, 38(5), 388-419

Jacobson, N. and Goering, P. (2006). Credibility and credibility work in knowledge transfer, Evidence and Policy, 2(2), 151-65.

Jones, M., & Kondylis, F. (2016). Lessons from a crowdsourcing failure. Retrieved from http://blogs.worldbank.org/impactevaluations/lessons-crowdsourcing-failure

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F. and B. Volkov (2009), Research on Evaluation Use: A Review of the Empirical Literature From 1986 to 2005, American Journal of Evaluation, vol. 30(3): 377-410

Karlan, D., & Appel, J. (2016a). *Failing in the field: What we can learn when field research is going wrong*. Princeton University Press: Woodstock, Oxfordshire.

Karlan, D., & Appel, J. (2016b). When the Juice Isn't Worth the Squeeze: NGOs refuse to participate in a beneficiary feedback experiment Retrieved from http://blogs.worldbank.org/impactevaluations/when-juice-isn-t-worth-squeeze-ngos-refuse-participate-beneficiary-feedback-experiment

Keller, J. M., & Savedoff, D. (2016). Improving Development Policy through Impact Evaluations: Then, Now, and Next? Retrieved from https://www.cgdev.org/blog/improving-development-policy-through-impact-evaluations-then-now-and-next

Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F. & Volkov, B. (2009) Research on Evaluation Use: A Review of the Empirical Literature From 1986 to 2005. *American Journal of Evaluation*, 30(3), 377-410.

Johnson, K. , Greenseid, L.O., Toal, S. A., King, J. A., Lawrenz, F., and Volkov, B. (2009) Research on Evaluation Use: A Review of the Empirical Literature From 1986 to 2005, *American Journal of Evaluatio*n, 30(3), 377-410.

Lavis, J.N., Robertson, D., Woodside, J.M., McLeod, C.B. and Abelson, J. (2003) How can research organisations more effectively transfer research knowledge to decision-makers? Milbank Quarterly, 81(2), 221-48.

Legovini, A., Di Maro, V., & Piza, C. (2015). Impact Evaluations Help Deliver Development Projects. *World Bank Policy Research Working Paper 7157*.

Lopez Avila, D. M. (2016). Implementing impact evaluations: trade-offs and compromises. Retrieved from http://blogs.3ieimpact.org/implementing-impact-evaluations-trade-offs-and-compromises/

Ludwig, J., Kling, J. R., and S. Mullainathan (2011) Mechanism Experiments and Policy Evaluation, Journal of Economic Perspectives, 25(3), 17-38.

MCC. (2012). MCC's First Impact Evaluations: Farmer Training Activities in Five Countries. *MCC Issue Brief*.

Mckenzie. (2016). Book Review: Failing in the Field – Karlan and Appel on what we can learn from things going wrong. Retrieved from http://blogs.worldbank.org/impactevaluations/book-review-failing-field-karlan-and-appel-what-we-can-learn-things-going-wrong

Mckenzie, D. (2012). Misadventures in Photographing Impact. Retrieved from http://blogs.worldbank.org/impactevaluations/misadventures-in-photographing-impact

Mckenzie, D. (2016a). Lessons from some of my evaluation failures: Part 1 of 2 Retrieved from http://blogs.worldbank.org/impactevaluations/lessons-some-my-evaluation-failures-part-1

Mckenzie, D. (2016b). Lessons from some of my evaluation failures: Part 2 of 2 Retrieved from http://blogs.worldbank.org/impactevaluations/lessons-some-my-evaluation-failures-part-2

McNulty, J. (2012). Symbolic use of evaluation in the international aid sector: arguments for critical reflection. Evidence and Policy, 8(4), 495-509.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L. (2017) Abandon Statistical Significance, http://www.stat.columbia.edu/~gelman/research/unpublished/abandon.pdf

MEASURE evaluation. (2015). Evaluation FAQ: How Much Will an Impact Evaluation Cost? *MEASURE evaluation brief November 2015*.

Mishra, A., & Cameron, D. (2014). Instsructions for Impact Evaluation Repository Screening. *International Initiative for Impact Evaluation (3ie), Washington DC*.

Nutley, S. M., Walter, I., & Davies, H. T. O. (2008). *Using Evidence: How Research Can Improve Public Services*. Bristol: The Policy Press.

Oliver, S., Cartwright, N., Bangpan, M., Dickson, K., Gough, D., Hargreaves, J., Pells, K., Roche, C., & Stewart, R. (2018) Stakeholder engagement for development impact and learning, CEDIL Inception paper.

Peersman, G., Guijt, I., & Pasanen, T. (2015). Evaluability assessment for impact evaluation: guidance, checklist and decision support. *A Methodslab Publication ODI*.

Priedeman Skiles, M., Hattori, A., & Curtis, S. L. (2014). *Impact Evaluations of Large-Scale Public Health Interventions Experiences from the Field*. USAID: MEASURE International Working Paper 14-17.

Pritchett, L., Samji, S., and J. Hammer (2013) It is All about MeE: Using Structural Experiential Learning ("e") to Crawl the Design Space, Working Paper 322, Centre for Global Development.

Puri, J., & Rathinam, F. (2017). Grants for real-world impact evaluations: what are we learning? *unpublished manuscript*.

Rossi, P.H., Lipsey, M.W., Freeman, H.E. (2004) Evaluation: A Systematic Approach, Thousand Oaks: Sage Publications Inc.

Samuels, F., & McPherson, F. (2010). Meeting the challenge of proving impact in Andhra Pradesh, India. *Journal of Development Effectiveness, 2*(4), 468-485.

Schünemann, H.J., Oxman, A.D., Vist, G.E., Higgins, J.P.T., Deeks, J.J., Glasziou, P., Guyatt, G.H. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0

(updated March 2011). The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA, US: Houghton, Mifflin and Company

Shapiro, A. K., & Shapiro, E. (1999). The Placebo: Is It Much Ado about Nothing? In H. Harrington (Ed.), The Placebo Effect: An Interdisciplinary Exploration. London: Harvard University Press.

Shula, M.L. & Cousins, J.B (1997) Evaluation Use: Theory, Research, and Practice Since 1986. *Evaluation Practice*, 18(3), 195-208.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., and B. Befani (2012). Broadening the Range of Designs and Methods for Impact Evaluation, DFID Working Paper No. 38, London.

Stickler, M. (2016). 5 Things USAID's Land Office Has Learned about Impact Evaluations. Retrieved from https://blog.usaid.gov/2016/05/mythbusting-5-things-you-should-know-about-impact-evaluations-at-usaid/

Sturdy, J. Aquino, S. and Molyneaux, J. (2014). Learning from evaluation at the Millennium Challenge Corporation, Journal of Development Effectiveness, 6(4), 436-450.

Turner, R. M., Bird, S. M. and Higgins, J. P. T. (2013) 'The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews', *PLoS ONE*, 8(3). doi: 10.1371/journal.pone.0059202.

Vaessen, J. (2017). Evaluability and why it is Important for Evaluators and Non-Evaluators. Retrieved from http://ieg.worldbankgroup.org/blog/evaluability

Vermeersch, C. (2012). "Oops! Did I just ruin this impact evaluation?" Top 5 of mistakes and how the new Impact Evaluation Toolkit can help. Retrieved from http://blogs.worldbank.org/impactevaluations/oops-did-i-just-ruin-this-impact-evaluation-top-5-of-mistakes-and-how-the-new-impact-evaluation-tool

Weiss, C. H. (1980). Knowledge Creep and Decision Accretion. *Knowledge: Creation, Diffusion, Utilization, 1*(3), 381-404.

Weiss, C.H. and Buculavas, M.J. (1980) Truth tests and utility tests: decision-makers' frames of reference for social science research, American Sociological Review, 45(2), 302-13.

White, H. (2009). Theory-Based Impact Evaluation: Principles and Practice, *Working Paper 14*. New Delhi: International Initiative for Impact Evaluation (3ie).

White, H. (2014). Ten things that can go wrong with randomised controlled trials. Retrieved from http://blogs.3ieimpact.org/ten-things-that-can-go-wrong-with-randomised-controlled-trials/

White, H. (2014) Is Independence Always a Good Thing? Retrieved from http://blogs.3ieimpact.org/is-independence-always-a-good-thing/

Wholey, J. (2004) Exploratory evaluation. In H.P. Hatry, J.S. Wholey, and K.E. Newcomer (Eds.), Handbook of practical program evaluation (2nd ed, 241-260). San Francisco, CA: Jossey-Bass.

Wood, B. (2015). Replication research promotes open discourse. *Evidence Matters* *http://blogs.3ieimpact.org/replication-research-promotes-open-discourse/*.

Wood, B. D. K., & Djimeu, E. (2014). Requiring fuel gauges: A pitch for justifying impact evaluation sample size assumptions. *Evidence Matters* *http://blogs.3ieimpact.org/requiring-fuel-gauges-a-pitch-for-justifying-impact-evaluation-sample-size-assumptions/*.

Contact Us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDILProgramme